Raport de analiză a impactului utilizării transcrierilor aproximative în vederea reantrenării sistemelor de RAV

Cristian Manolache, Alexandru-Lucian Georgescu,

Horia Cucu, Dragoș Burileanu, Corneliu Burileanu

1. Introducere

După o primă discuție sumară despre seturile de date utilizate în cadrul acestui raport, vom aborda următoarele subiecte: (i) îmbunătățirea metodei de utilizare a transcrierilor aproximative pentru generarea de adnotări pentru seturi de date de vorbire şi (ii) aplicarea metodei pe două noi seturi de date brute: corpusul CoBiLIRo şi corpusul CDep, urmată de evaluarea impactului seturilor de date nou adnotate asupra sistemului de RAV inițial.

Acest raport de analiză face referire în multiple rânduri la raportul ştiințific şi tehnic aferent proiectului component TADARAV [Georgescu, 2020b] disponibil online pe pagina web a acestui proiect de cercetare[1].

1.1 Seturi de date

Seturile de date la care se face referire în cadrul acestui raport sunt prezentate în Tabelul 1 şi Tabelul 2. Mai multe detalii despre aceste seturi de date puteți găsi în raportul etapei 2019 a proiectului TADARAV [Georgescu, 2020b], secțiunea 2.1. Este de menționat faptul că seturile de date de evaluare RSC-eval [Georgescu, 2020a] şi SSC-eval1 au fost actualizate în cursul anului 2020. Toate comparațiile rezultatelor de recunoaaştere automată a vorbirii (RAV) din acest raport cu rezultate publicate anterior anului 2020 trebuie să țină cont de îmbunătățirea artificială de performanță provenită din corectarea acestor seturi de date de evaluare. Mai multe detalii în [Georgescu, 2020b], secțiunea 2.1.

2. Îmbunătăţirea soluţiei de aliniere a transcrierilor aproximative cu semnalul de vorbire

2.1 Metoda de bază

Metoda de bază de filtrare şi aliniere a transcrierilor aproximative cu semnalul de vorbire a fost introdusă, prezentată şi evaluată în activitatea 2.11 din etapa 2019 a proiectului (vezi raport TADARAV 2019 [Georgescu, 2019a], secțiunea 2.2). În cele ce urmează este prezentată o descriere sumară a acestei metode.

[1] Proiectul de cercetare TADARAV: https://tadarav.speed.pub.ro

Obiectivul principal al metodei de bază de aliniere este obținerea adnotărilor precise pentru o parte a unui corpus de vorbire în mod automat. Noul corpus obținut va fi folosit pentru reantrenarea sistemelor RAV existente crescând astfel variabilitatea modelului acustic, ceea ce ar trebui să implice mai departe o îmbunătățire în acuratețea transcrierilor. Metoda de adnotare se bazează pe folosirea unui corpus de vorbire brut împreună cu transcrieri aproximative şi transcrierile produse de un sistem RAV inițial. Cele două seturi de transcrieri ale corpusului de vorbire sunt aliniate folosind distanța Levenstein. În urma alinierii, considerăm părțile identice dintre cele două transcrieri ca fiind corecte şi le selectăm ca parte a noului corpus.

Părțile de vorbire şi de transcrieri aproximative au fost extrase din mediul online mass-media (ştiri, interviuri, rapoarte). Deşi mediul online mass-media este o sursă bogată de vorbire şi text, transcrierile aproximative extrase sunt diferite de cele generate de sistemul RAV în ceea ce priveşte formatul. Transcrierile aproximative conțin titluri, nume proprii, semne de punctuație, numere, abrevieri etc., pe când transcrierile RAV sunt practic secvențe de cuvinte cu litere mici însoțite de etichete de timp. Ca şi exemplu:

Tabelul 1. *Seturile de vorbire adnotată folosite pentru antrenarea şi evaluarea sistemelor de RAV şi seturile de vorbire adnotată obținute în etapa anterioară (2/2019)*

Setul de date	Subset	Durată
Spontaneous Speech Corpus (SSC)	SSC-train1+2	130h, 44m
	SSC-train3-trans-v4	41h, 00m
	SSC-train4-trans-v4	250h, 10m
	SSC-eval1	3h, 29m
	SSC-eval2	1h, 31m
Read Speech Corpus (RSC)	RSC-train	94h, 46m
	RSC-eval	5h, 29m
Contemporary Romanian Language (CoRoLa)	n/a	85h, 11m
Camera Deputaților (CDep)	CDep-eval	5h, 00m

Tabelul 2. *Seturi de date de vorbire neadnotată (+ transcrieri aproximative) utilizate ca date de intrare pentru cele trei metode de adnotare automată.*

Setul de date	Sursa	Durată	Transcrieri aproximative	Număr de vorbitori
CDep-raw	Parlamentul României	3,510h, 13m	25.1M cuvinte	~2,500
CoBiLiRo-raw	Alma[2]	19h, 29m	127.5k cuvinte	n/a
	VBarbu[3]	7h, 30m	80.8k cuvinte	
	GCVC[4]	16h, 49m	324.5k cuvinte	
	Ro100[5]	25h, 35m	181.8k cuvinte	

[2] Seria de emisiuni Alma Mater Iassiensis
[3] Interviu cu prof. Viorel Barbu
[4] Seria de emisiuni Ghici cine mai vine la cină
[5] Seria de emisiuni România 100: Iaşul în arcul timpului

- Transcriere aproximativă: Bărbatul de 36 de ani povesteşte că muncise toată noaptea.
- Transcriere RAV: bărbatul(3.71, 4.02) de(4.02, 4.87) treizeci(4.87, 5.11) şi(5.11, 5.74) şase(5.74, 6.02) de(6.02, 6.35) ani(6.35, 6.71) povesteşte(6.71, 6.94) că(6.94, 7.58) muncise(7.58, 7.89) toată(7.89, 8.16) noaptea(8.16, 8.32)

Astfel, transcrierile aproximative trebuie supuse unor operații de procesare de text, anume: restaurare de diacritice, înlocuirea URL-urilor şi adreselor de email cu forma lor vorbită, înlocuirea numerelor cu text, expandarea abrevierilor, înlocuirea caracterelor speciale, transformarea literelor mari în litere mici.

După procesul de aliniere, segmentele de text obținute sunt filtrate pe baza anumitor criterii cu scopul de a înlătura segmente foarte mici (în ceea ce priveşte durata vorbirii şi numărul de caractere) şi segmente ce conțin zone lungi de linişte (durata dintre cuvinte consecutive). În final, procesele de aliniere şi filtrare generează un set de transcrieri aliniate însoțite de etichete de timp. Etichetele de timp vor fi folosite mai departe pentru tăierea segmentelor audio corespunzătoare din materialele audio. Astfel, noul corpus de vorbire adnotată va conține transcrieri aliniate şi segmentele audio corespunzătoare, ce pot fi folosite mai departe pentru reantrenarea sistemelor RAV.

2.2 Metoda propusă

Metoda descrisă anterior este foarte strictă în ceea ce priveşte corectitudinea transcrierilor aliniate, în sensul că sunt considerate a fi corecte doar secvențele de cuvinte unde transcrierile aproximative şi transcrierile RAV se potrivesc perfect. Acest fapt duce la selectarea multor secvențe scurte (*i.e.* ce cuprind puține cuvinte). În urma unei analize manuale a transcrierilor s-a constatat că există multe secvențe de câteva cuvinte pentru care procesul de aliniere a eşuat şi care se află între secvențe lungi de cuvinte care s-au aliniat cu succes. Câteva astfel de exemple sunt prezentate în Tabelul 3.

Aceste secvențe nu au fost aliniate din cauza mai multor motive precum zgomot de fundal în materialul audio, cuvinte care nu se regăsesc în vocabularul sistemului RAV (de regulă nume proprii), cuvinte adiționale în fişierul audio care nu sunt pronunțate în materialul audio. Cel de-al treilea caz se întâlneşte de obicei în interviuri, unde transcrierea de pe website conține text adițional pentru a preciza cine vorbeşte.

În urma unei analize a 50 de astfel de situații, am ajuns la concluzia că dacă am lua în considerare secvențe de mai puțin de 6 cuvinte pentru care procesul de aliniere a eşuat, în aproximativ 68% din cazuri transcrierea aproximativă este corectă, pe când transcrierea RAV este incorectă. Astfel, în aceste cazuri am decis să unim cele 3 secvențe de cuvinte. Adăugând aceste secvențe de text sperăm să generăm un corpus de vorbire mai mare ce va fi folositor pentru reantrenarea sistemelor RAV. Noi credem că aceste secvențe vor fi mai valoroase pentru reantrenare, deoarece sistemul RAV inițial nu a reuşit să le transcrie. Desigur, există şi o parte negativă la acest compromis: 32% din secvențele adăugate sunt incorecte şi pot afecta în mod negativ procesul de antrenare.

***Tabelul 3**. Exemple de secvențe de cuvinte pentru care procesul de aliniere a eşuat şi care se află între secvențe lungi de cuvinte care s-au aliniat cu succes. Secvențele pentru care procesul de aliniere a eşuat sunt marcate cu roşu/ verde având în vedere dacă transcrierea aproximativă este incorectă/corectă.*

Nr.	Secvența anterioară (aliniată)	Secvență nealiniată	Secvența următoare (aliniată)	Observații
1	… se îndreaptă către susținătorii săi	o dronă	filmează evenimentul ...	confuzat cu "un pod" de către sistemul RAV din cauza zgomotului de fundal
2	… în zona	Egiptului	unde există ...	confuzat cu "edituri" de către sistemul RAV deoarece acest cuvânt nu se află în vocabular
3	… a obținut fondurile	administrator bloc	la doi ani …	secvența de text nu este pronunțată în materialul audio, scopul acesteia în transcrierea aproximativă este aceea de a specifica cine vorbeşte

2.3 Evaluarea metodelor

O metodă de adnotare automată poate fi evaluată în funcție de dimensiunea corpusului nou generat şi în funcție de calitatea transcrierilor, exprimată în funcție de rata de eroare la nivel de cuvânt (WER). Dimensiunea corpusului nou generat se doreşte a fi pe cat de mare posibilă. Desigur, limita superioară este dimensiunea corpusului audio brut (materialele audio). Astfel, o metrică mai potrivită este eficacitatea metodei, definită ca procentajul de material audio ce a fost selectat ca făcând parte din corpusul nou generat şi pentru care metoda a generat transcrieri corecte. Calitatea adnotării exprimată în rata de eroare la nivel de cuvânt (WER) şi/sau rata de eroare la nivel de caracter (ChER) nu poate fi măsurată pentru metoda de bază din cauza absenței unei referințe pentru transcrierile aproximative. Totuşi, evaluarea se poate face pe alt criteriu care se bazează pe această metrică. O adnotare de calitate înaltă ar trebui să ducă în mod normal la o performanță mai bună a sistemului RAV reantrenat pe noul corpus de vorbire generat.

2.4 Rezultate alinieri

Procedura de aliniere şi filtrare prezentată în secțiunea 2.1 a fost aplicată pe seturile de date brute SSC-train3-raw şi SSC-train4-raw. În etapa 2/2019 am folosit ca punct de pornire transcrierile acestor două seturi de date generate cu un sistem RAV inițial mai slab, şi anume cel prezentat în [Georgescu, 2017]. Au rezultat astfel seturile de date de vorbire adnotată SSC-train3-trans-v3 şi SSC-train4-trans-v3. Dimensiunile acestora, exprimate în număr de ore de vorbire şi eficiența procesului de adnotare automată, exprimată sub forma procentului de date brute ce au putut fi adnotate, raportat la dimensiunea datelor brute sunt prezentate în Tabelul 4.

Aceeaşi metodă de aliniere şi filtrare a fost aplicată în cadrul activității curente folosind ca punct de pornire transcrierile acestor două seturi de date generate cu un sistem RAV îmbunătățit, şi anume cel prezentat în [Georgescu, 2019b]. A rezultat astfel versiunea v4 a acestor seturi de date adnotate, versiune ce cuprinde mai multe ore de vorbire, conform Tabelului 4.

În urma aplicării procedurii de aliniere şi filtrare prezentată în secțiunea 2.2 pe seturile de date brute SSC-train3-raw şi SSC-train4-raw, au fost obținute versiunile v5 ale seturilor de date adnotate. În acest caz a fost utilizat pentru transcriere tot sistemul de RAV îmbunătățit prezentat în [Georgescu, 2019b].

Aşa cum era de aşteptat, seturile de date v4 sunt mai mari (cu aproximativ 9%) decât seturile de date v3. Eficiența alinierii a crescut de la 27.4% la 30.0% pentru primul set de date, respectiv de la 29.4% la 32.2% pe cel de-al doilea set de date. Ne aşteptăm ca această creştere a seturilor de date să contribuie pozitiv la reantrenarea sistemului de RAV inițial, întrucât creşterea se bazează în totalitate pe nişte transcrieri RAV inițiale mai precise.

De asemenea, seturile de date v5 sunt mai mari (cu aproximativ 12%) decât seturile de date v4. Eficiența alinierii a crescut de la 30.0% la 33.7% pentru primul set de date, respectiv de la 32.3% la 36.2% pe cel de-al doilea set de date. Urmează să vedem, în secțiunea următoare, dacă aceasta creştere a seturilor de date contribuie pozitiv sau negativ la reantrenarea sistemului de RAV inițial.

Tabelul 4. *Statistici pentru seturile de date SSC-train3-trans-v4 şi SSC-train4-trans-v4*

Set de date	Durată [# ore]	Eficiență aliniere [% ore]
SSC-train3-trans-v3	37,5	27,4%
SSC-train3-trans-v4	41,0	30,0%
SSC-train3-trans-v5	46,1	33,7%
SSC-train4-trans-v3	228,8	29,4%
SSC-train4-trans-v4	250,2	32,2%
SSC-train4-trans-v5	281,6	36,2%

2.5 Sisteme RAV ce utilizează noile corpusuri obţinute

Sistemele RAV obținute în urma reantrenării cu noile seturi de date sunt prezentate în Tabelul 5, alături de sistemul RAV inițial (cel ce a fost utilizat pentru transcrierea datelor brute în anul 2020). Comparativ cu raportul precedent, evaluarea s-a realizat pe seturile de evaluare corectate RSC-eval-v2 şi SSC-eval-v2 (vezi secțiunea 1.1 pentru mai multe detalii).

***Tabelul 5**. Performanța sistemelor RAV după reantrenare*

Sistem RAV	Set antrenare	WER [%]		Îmbunătățire relativă a WER [%]	
		RSC-eval	SSC-eval1	RSC-eval	SSC-eval1
Sistem RAV inițial	RSC-train + SSC-train	1.88	14.96	n/a	n/a
RAV reantrenat 2019 (metoda de bază)	RSC-train + SSC-train + SSC-train3-trans-v3 + SSC-train4-trans-v3	1.83	11.50	2.66%	23.1%
RAV reantrenat 2020 (metoda de bază)	RSC-train + SSC-train + SSC-train3-trans-v4 + SSC-train4-trans-v4	1.83	11.04	2.66%	26.2%
RAV reantrenat 2020 (metoda propusă)	RSC-train + SSC-train + SSC-train3-trans-v5 + SSC-train4-trans-v5	1.83	11.13	2.66%	25.6%

În ceea ce priveşte rezultatele de transcriere de vorbire continuă (setul RSC-eval), observăm că indiferent de metoda de aliniere şi filtrare utilizată şi indiferent de calitatea transcrierilor inițiale, sistemele RAV reantrenate pe seturile inițiale şi seturile nou adnotate prezintă aceeaşi îmbunătățire față de sistemul RAV inițial: WER de 1.83% vs. 1.88%.

Pentru vorbire spontană, metoda de bază aplicată anul trecut, folosind transcrieri inițiale de calitate mai slabă, conduce la obținerea unei îmbunătățiri relative de WER de aproximativ 23%. Aceeaşi metodă, aplicată în acest an, pornind de la transcrieri inițiale mai precise, conduce la obținerea unei îmbunătățiri relative de WER de aproximativ 26%. Se observă astfel importanța calității transcrierilor inițiale.

Metoda nou propusă în acest an conduce la obținerea unei îmbunătățiri relative de WER de aproximativ 25% față de sistemul RAV inițial. Se observă astfel că propunerea de a utiliza în antrenarea RAV şi a secvențelor de câteva cuvinte (maxim 6) pentru care procesul de aliniere a eşuat şi care se află între secvențe lungi de cuvinte care s-au aliniat cu succes, nu conduce la rezultate RAV mai bune. Din nou rezultă faptul că este extrem de importantă calitatea transcrierilor pentru datele de antrenare, în detrimentul cantității acestor date.

3. Aplicarea metodei pe setul de date CoBiLiRo-raw

3.1 Analiza inițială a setului de date CoBiLiRo-raw

Metoda de bază, descrisă în secțiunea 2.1, a fost aplicată şi pe setul de date CoBiLiRo-raw (vezi Tabelul 2), cu scopul îmbunătățirii sistemului RAV. Această metodă a fost preferată față de metoda propusă în secțiunea anterioară deoarece cea din urmă a fost evaluată cu rezultate mai slabe. În primă fază, s-a aplicat metoda pe doar 2 fişiere din setul de date, în vederea analizării eventualelor probleme. Rezultatele alinierii se pot observa în Tabelul 6, unde sunt prezentate numărul de cuvinte aliniate, numărul de cuvinte ce se regăsesc în transcrierea aproximativă, numărul de cuvinte ce se regăsesc în transcrierea RAV, precum şi procentul de cuvinte aliniate raportat la numărul de cuvinte din transcrierea aproximativă.

***Tabelul 6**. Transcrieri RAV obținute cu sistem RAV baseline [Georgescu, 2019b]*

Fişierul audio	Nr. cuvinte aliniate	Nr. cuvinte transcrieri aproximative	Nr. cuvinte transcrieri RAV	Aliniere [% cuvinte]
alma24	3.696	5.151	5.985	71,75%
interviu2	2.065	18.095	11.225	18,40%

După analiza rezultatelor din Tabelul 6, precum şi a fişierelor în cauză, s-au constatat următoarele: (i) în cazul fişierului alma24, transcrierea RAV conține mai multe cuvinte decât cea aproximativă, deoarece în fişierul audio sunt pronunțate mai multe cuvinte decât există în transcrierea aproximativă, (ii) în cazul fişierului interviu2, transcrierea aproximativă conține foarte mult text care nu este pronunțat în fişierul audio. Există reformulări, adăugiri, precum şi precizări suplimentare.

În Tabelele 7 şi 8 sunt prezentate câteva exemple de probleme întâlnite pentru cele 2 seturi de date. În Tabelul 7, unde sunt prezentate exemple de text lipsă pentru setul de date alma24, se poate observa pe prima coloană text care este întâlnit în transcrierea aproximativă dar care este şi pronunțat în fişierul audio, urmat de eticheta de timp la care se încheie pronunția textului respectiv pe coloana a 2-a. Acest text este situat anterior față de cel de pe coloana a 3-a, text care este pronunțat în fişierul audio dar nu se regăseşte în transcrierea aproximativă. Pe coloanele 4 şi 5 din Tabelul 7 se regăseşte text similar cu cel din primele 2 coloane cu excepția faptului că acest text este situat ulterior față de cel de pe coloana 3, iar eticheta de timp marchează momentul de timp la care se începe pronunția textului. În final, pe ultima coloană se regăseşte diferența dintre cele 2 etichete de timp, care marchează durata de timp pentru care lipseşte transcrierea de pe coloana 3. Analizând exemplele din Tabelul 7, se poate observa că există porțiuni din fişierul audio de până la aproximativ 20s pentru care nu există transcriere aproximativă.

În Tabelul 8 sunt prezentate exemple de text care se regăseşte în transcrierea aproximativă dar nu este pronunțat în fişierul audio. După cum se poate observa, există atât secvențe scurte de text nepronunțat cât şi secvențe mai lungi.

***Tabelul 7**. Analiză probleme set de date alma24*

Context anterior (există atât în audio cât şi în transcrierea aproximativă)	Etichetă timp context anterior	Există în fişierul audio, dar nu există în transcrierea aproximativă	Context ulterior (există atât în audio cât şi în transcrierea aproximativă)	Etichetă timp context ulterior	Delta timp fără transcriere [s]
de când realizăm această emisiune	29,28	universală de dar în colaborare cu tvr iaşi sau tvr iaşi doar cu universitatea pentru a marca împlinirea a o sută cincizeci de ani de la crearea universității moderne a fost trimisă de variate până habar nici despre istoria universității emisiuni despre facultății şi cu diverse personalități	spuneam într-o discuție anterioară	49,77	20,49
nouăzeci de ani	133,26	pentru ştiu că sunt program complex	am publicat	142,8	9,54
eu	311,43	sunt un pic mai sus	cred că	315,45	4,02
Curtea Constituțională	413,76	nu o să vă întreb despre legea pensiilor discutăm despre facultatea de drept	Domnule profesor	420	6,24

***Tabelul 8**. Analiză fişier interviu2*

Context anterior (există atât în audio cât şi în transcrierea aproximativă)	Etichetă timp context anterior	Există în transcrierea aproximativă, dar nu există în fişierul audio
universitatea din freiburg	110	în toate domeniile şi la toate palierele vasta sa activitate de construcție a relației dintre iaşi şi freiburg a început
foarte importante	165,99	pentru noi profesorii şi deplasări în germania pe atunci relația cu freiburg
realizările lui miron	205	poate cea mai importantă
îşi pot imagina	218,34	ce şansă enormă însemna pentru acei oameni simpla posibilitate de a ieşi fie şi pentru perioade scurte din închisoarea generalizată care era românia comunistă da acest lucru a însemnat ceva excepțional
îmi amintesc	223	că prin o mie nouă sute optzeci şi cinci a avut loc în germania un mare
o proprietate	337,08	unde au construit o casă unde continuă să îşi primească prietenii în plus au ctitorit acolo o frumoasă bisericuță
înmormântat	341	în cimititul din imediata apropriere decizia de a-şi găsi odihna la malul mării nu poate fi desprinsă cred de simbolistica exilului atât de apropiată fostului mare exilat de altfel paul miron a şi scris o piesă de teatru despre ovidiu alt mare exilat dar pe țărmurile pontului euxin

După analiza acestor probleme, s-a încercat o actualizare a modelelor de limba folosind transcrierile aproximative ale celor 2 seturi de date ca text adițional, obținând rezultatele din Tabelul 9.

Comparând rezultatele din Tabelul 9 cu cele inițiale din Tabelul 6, se poate observa o creştere a numărului de cuvinte aliniate pentru setul alma24, dar o scădere drastică pentru setul interviu2. După realizarea unor seturi de experimente, în care s-au aliniat anumite părți de text din transcrierea aproximativă cu partea corespunzătoare a transcrierii RAV a setului interviu2 (prima jumătate a transcrierii aproximative cu prima jumătate a transcrierii RAV, a doua jumătate, primele 1000 de cuvinte, primele 2000 de cuvinte, … , primele 8000 de cuvinte) s-a constatat că modulul de aliniere din aplicația JavaNLP2 (NISTAlign, prezentat în raportul anterior, secțiunea 2.2.3) întâlneşte probleme când primeşte la intrare cantități mari de text. Ca soluție pentru această problemă, partea de aliniere a fost mutată în exteriorul aplicației JavaNLP2 şi executată cu toolkit-ul sclite, urmând însă ca operațiile de filtrare să fie executate tot de javaNLP2. Rezultatele actualizate sunt prezentate în Tabelul 10.

Analizând rezultatele din Tabelul 10, putem observa o uşoară creştere a numărului de cuvinte aliniate în cazul setului alma24 comparativ cu rezultatele din Tabelul 9. De asemenea, pentru setul interviu2, problema generată de către aliniatorul NISTAlign din toolkit-ul CMU Sphinx a fost rezolvată folosind modulul alternativ de aliniere din toolkit-ul sclite. Deşi această problemă a fost rezolvată, procentul de cuvinte aliniate rămâne relativ mic.

***Tabelul 9**. Transcrieri RAV obținute cu sistem cu model de limbă actualizat*

Set de date	Nr. cuvinte aliniate	Nr. cuvinte transcrieri aproximative	Nr. cuvinte transcrieri RAV	Aliniere [% cuvinte]
alma24	4.137	5.151	5.621	80,31%
interviu2	33	18.095	8.830	0,18%

***Tabelul 10**. Transcrieri RAV obținute cu sistem cu model de limbă actualizat + aliniere cu sclite*

Set de date	Nr. cuvinte aliniate	Nr. cuvinte transcrieri aproximative	Nr. cuvinte transcrieri RAV	Aliniere [% cuvinte] din transcrierea aproximativă	Aliniere [% cuvinte] din transcrierea RAV
alma24	4.157	5.151	5.621	80,70%	73,95%
interviu2	3.811	18.095	8.830	21,06%	43,16%

3.2 Aplicarea metodei de bază set de date CoBiLiRo-raw

În urma testării metodei de bază pe cele 2 fişiere de test, metoda de bază a fost aplicată pe întreg setul de date CoBiLiRo-raw, cu mențiunea că procesul de aliniere a fost efectuat de toolkit-ul sclite şi nu de CMU Sphinx (NISTAlign) pentru a evita eventualele probleme similare cu cele întâlnite la setul de date interviu2, menționate în subcapitolul anterior. Sistemul RAV folosit pentru transcrierea întregului set de date CoBiLiRo-raw a fost sistemul RAV baseline [Georgescu, 2019b]. Sistemul RAV cu model de limbă actualizat putea conduce, în final, la obținerea unui set de date adnotat mai mare, dar fără garanția că acel text ar fi fost era corect. Din acest motiv sistemul RAV cu model de limbă actualizat nu a fost folosit. În urma aplicării metodei de bază s-a obținut setul de date CoBiLiRo-trans-v4 ce conține secvențele de cuvinte aliniate pentru fiecare subset de date din CoBiLiRo-raw, împreună cu etichete de timp pentru fiecare cuvânt din secvență ce indică poziția şi durata lor în fişierul audio corespunzător. Detalii legate de setul de date CoBiLiRo-trans-v4 se găsesc în Tabelul 6.

Sistemul RAV reantrenat cu seturile de date inițiale şi setul de date nou adnotat (CoBiLiRo-trans-v4) a obținut o rată de eroare la nivel de cuvânt de 1.8% pe setul RSC-eval ce conține vorbire citită, respectiv 14.0% pe setul SSC-eval1 ce conține vorbire spontană. Astfel, adăugarea setului CoBiLiRo-trans-v4 de 31 ore la setul inițial de antrenare de 225 ore, a condus la o îmbunătățire relativă față de sistemul inițial cu 5.2% pe vorbire citită, respectiv 6.6% pe vorbire spontană.

4. Aplicarea metodei pentru setul de date CDep-raw

Metoda de bază prezentată în secțiunea 2.1 a fost aplicată şi pe setul de date CDep-raw (vezi Tabelul 2). Fişierele audio din setul de date CDep-raw au fost segmentate la 1 minut, obținând setul de date CDep-raw-1min. Setul de fişiere audio CDep-raw-1min a fost transcris utilizând sistemul RAV care a obținut cele mai bune rezultate în analiza făcută în 2019 [Georgescu, 2021]. A rezultat astfel setul de transcrieri RAV necesare pentru efectuarea procesului de aliniere din metoda de bază. După transcrierea fişierelor audio de 1 minut, 300 dintre acestea au fost separate, împreună cu transcrierile lor, în vederea formării unui set de date de evaluare CDep-eval. Transcrierile RAV ale celor 300 de fişiere audio au fost apoi aliniate cu transcrierile aproximative. Deoarece secvențele aliniate nu conțin întreg textul vorbit din fişierul audio a fost necesară intervenția manuală prin analiza şi corectarea fişierelor text în vederea obținerii unor transcrieri de referință ce pot fi folosite mai departe pentru evaluarea viitoarelor sisteme RAV. Setul de transcrieri obținute, exceptând cele din setul CDep-eval, au fost alipite astfel încât timpii să se refere la fişierul audio original şi nu la fişierul audio tăiat. După efectuarea procesului de aliniere între transcrierile RAV obținute după alipire şi transcrierile aproximative CDep s-a obținut setul de date CDep-trans-v4, ce conține 21M de cuvinte şi 878.8 ore de vorbire (84,2% din cuvinte, respectiv 25% din numărul de ore). Analizând rezultatele se poate observa o posibilă neconcordanță între numărul mare de cuvinte aliniate şi numărul mic de ore, ceea ce a dus spre efectuarea unor analize suplimentare. Aceste analize au constat în evaluarea fişierelor în funcție de energie şi putere. Prin efectuarea acestor analize, ne-am aşteptat să întâlnim un număr mare de fişiere cu un nivel de putere sub un anumit prag. Ne-am fi aşteptat să găsim fişiere ce conțin linişte pe întreaga sau pe majoritatea duratei fişierului audio, ceea ce ar fi justificat numărul mic de ore raportat la numărul mare de cuvinte aliniate. După efectuarea evaluărilor şi a unei analize manuale a mai multor fişiere audio, am observat că există şi fişiere cu nivel relativ mic de putere ce conțin în mare parte vorbire decât linişte, însă nu am putut obține încă niște rezultate cantitative care să justifice neconcordanța menționată mai sus. Această problemă urmează să mai fie analizată în cadrul etapei 2021 a proiectului.

Sistemul RAV care a fost utilizat pentru generarea transcrierilor setului CDep-raw a fost reantrenat cu seturile de date inițiale şi noul set adnotat CDep-trans-v4. Sistemul RAV reantrenat a fost evaluat pe seturile de evaluare RSC-eval şi SSC-eval1 şi SSC-eval2, iar rezultatele sunt prezentate în Tabelul 11 alături de cele ale sistemului RAV inițial.

Tabelul 11. *Sistem RAV baseline vs sistem RAV*

Sistem RAV	**Seturi de antrenare**	**WER [%]**		
		RSC-eval	**SSC-eval1**	**SSC-eval2**
Sistem RAV inițial [Georgescu, 2021]	RSC-train + SSC-train1+2 + SSC-train3+4-trans-v4	1.8	11.0	14.0
Sistem RAV reantrenat	RSC-train + SSC-train1+2 + SSC-train3+4-trans-v4 + CDep-trans-v4	1.7	12.3	15.4

Analizând rezultatele din Tabelul 11, putem observa o uşoară scădere a WER de la 1.8% la 1.7% în cazul vorbirii citite (RSC-eval) şi o degradare de aproximativ 10% a performanței pentru vorbire spontană (seturile SSC-eval1 şi SSC-eval2). Din aceste rezultate putem trage concluzia că, în ciuda adăugării setului de date CDep-trans-v4 (879h) la antrenarea modelului acustic, noul sistem RAV prezintă o performanță mai scăzută per total față de sistemul inițial. O posibilă explicație pentru această performanță mai scăzută este faptul că prelegerile din şedințele camerei deputaților, ce se regăsesc în setul CDep-trans-v4, conțin în mare parte vorbire citită (similară cu cea din RSC şi diferită față de SSC) şi cuvinte dintr-un domeniu diferit față de cel întâlnit în seturile de evaluare (ştiri, interviuri, radio).

Bibliografie

[Georgescu, 2017] A.-L. Georgescu, H. Cucu, C. Burileanu, "*SpeeD's DNN Approach to Romanian Speech Recognition,*" in Proc. 9th Conference on Speech Technology and Human-Computer Dialogue (SpeD), 8p, 2017.

[Georgescu, 2018] A.-L. Georgescu, H. Cucu, "Automatic annotation of speech corpora using complementary GMM and DNN acoustic models," In Proc. TSP 2018, pp. 1-4.

[Georgescu, 2019a] A.-L. Georgescu, C. Manolache, G. Pop, D. Oneață, H. Cucu, D. Burileanu, C. Burileanu, *"Proiect component TADARAV: Raport ştiințific şi tehnic în extenso 2019".*

[Georgescu, 2019b] A.-L. Georgescu, H. Cucu, C. Burileanu, *"Kaldi-based DNN architectures for speech recognition in Romanian,"* in the Proceedings of the 10th Conference on Speech Technology and Human-Computer Dialogue (SpeD), 2019, Timişoara, Romania.

[Georgescu, 2020a] A.-L. Georgescu, H. Cucu, A. Buzo, C. Burileanu, "*RSC: A Romanian Read Speech Corpus for Automatic Speech Recognition,*" in the Proceedings of The 12th Language Resources and Evaluation Conference (LREC), pp. 6606-6612, 2020, Marseille, France.

[Georgescu, 2020b]A.-L. Georgescu, A. Caranica, C. Manolache, G. Pop, D. Oneață, H. Cucu, C. Burileanu, D. Burileanu, *„Proiect component TADARAV: Raport ştiințific şi tehnic în extenso 2020".*

[Georgescu, 2021] A.-L. Georgescu, C. Manolache, D. Oneață, H. Cucu, C. Burileanu, "*Data-filtering methods for self-training of automatic speech recognition systems,*" in the Proceedings of the IEEE Spoken Language Technology Workshop (SLT), Virtual, 2021.